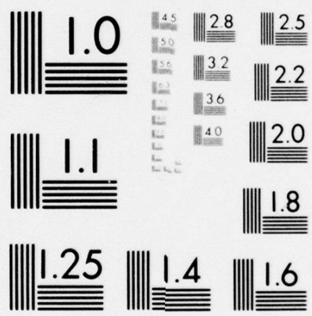
AD-A037 841 CENTER FOR NAVAL ANALYSES ARLINGTON VA MARINE CORPS --ETC F/G 12/1
THE EXACT DISTRIBUTION OF A SMIRNOV STATISTIC.(U)
FEB 77 E A PARENT, J K TYSON N00014-76-C-0001
UNCLASSIFIED CRC-317 NL

1 OF 1
ADA037841

END
DATE
FILMED
4-77


1.0
4.5
5.0
5.6
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.1
1.8
1.25
1.4
1.6

CRC 317

(12)
NW

THE EXACT DISTRIBUTION OF A SMIRNOV STATISTIC

CENTER FOR NAVAL ANALYSES
1401 Wilson Boulevard
Arlington, Virginia 22209
Marine Corps Operations Analysis Group

By: E.A. Parent, James K. Tyson

February 1977

Prepared for:

OFFICE OF NAVAL RESEARCH
Department of the Navy
Arlington, Virginia 22217

DEPUTY CHIEF OF STAFF (RD&S)
Headquarters, Marine Corps
Washington, D.C. 20380

02 031700.00

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

(14)

1. REPORT NUMBER: CRC-317
2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER

(6)

4. TITLE (and Subtitle): The Exact Distribution of a Smirnov Statistic.
5. TYPE OF REPORT & PERIOD COVERED
6. PERFORMING ORG. REPORT NUMBER

(10)

7. AUTHOR(s): E. A. Parent, James K. Tyson

(15)

8. CONTRACT OR GRANT NUMBER(s): N00014-76-C-0001

9. PERFORMING ORGANIZATION NAME AND ADDRESS:
Center for Naval Analyses
1401 Wilson Boulevard
Arlington, Virginia 22209

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS:
Deputy Chief of Staff (RD&S)
Headquarters, Marine Corps
Washington, D.C. 20380

(11)

12. REPORT DATE: February 1977
13. NUMBER OF PAGES: 12

(9)

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Research contribution

15. SECURITY CLASS. (of this report): Unclassified

(12) 17p.

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Cleared for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

(9)

18. SUPPLEMENTARY NOTES

This Research Contribution does not necessarily represent the opinion of the Department of the Navy.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

distribution theory, empirical distribution, statistical analysis, statistical tests, Smirnov

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The exact distribution is given for the number of times the vertical steps of an empirical distribution cross the underlying theoretical distribution. This statistic was used for Smirnov [1] to derive limiting distributions of statistics used in Kolmogorov-Smirnov tests, and may be used as a goodness-of-fit test in its own right.

100

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

401327

4B

Center for Naval Analyses

1401 Wilson Boulevard   Arlington, Virginia 22209   703/524-9400

An Equal Opportunity Employer

an affiliate of the University of Rochester

10 March 1977

MEMORANDUM FOR DISTRIBUTION LIST

Subj: Center for Naval Analyses Research Contribution 317

Encl: (1) CRC 317, "The Exact Distribution of a Smirnov Statistic," by E.A.Parent and James K. Tyson February 1977

1. Enclosure (1) is forwarded as a matter of possible interest.

2. Research Contributions are distributed for their potential value in other studies or analyses. They do not necessarily represent the opinion of the Department of the Navy.

ROBERT J. CORN
Director
Marine Corps Operations
Analysis Group

Subj: Center for Naval Analyses Research Contribution 317

Distribution List:

Commander, Operational Test and Evaluation Force
Deputy Commander, Operational Test and Evaluation Force, Pacific
Assistant Secretary of the Navy (Research and Development)
Office of Program Appraisal, Department of the Navy
Chief of Naval Research
DC/S for Research, Development & Studies, HqMC
National Defense University
Armed Forces Staff College
Annapolis Laboratory, David Taylor Naval Ship R&D Center
Carderock Laboratory, David Taylor Naval Ship R&D Center
New London Laboratory, Naval Undersea Center
AUTEC Detachment, Naval Undersea Center
San Diego Detachment, Naval Undersea Center
Newport Laboratory, Naval Undersea Center
Naval Surface Weapons Center
Naval Research Laboratory
Fleet Material Support Office
U.S. Naval Academy (Nimitz Library)
Naval Postgraduate School
Naval War College
Marine Corps Development and Education Command
CNO: Op-09BH and Op-96
Director, Defense Research and Engineering
Director, Planning and Evaluation (OSD)(2 copies)
Defense Documentation Center (12 copies)
Institute for Defense Analyses
The Rand Corporation
University of Rochester (2 copies)

TABLE OF CONTENTS

1. INTRODUCTION

Smirnov [1] defined a statistic V_n as the number of times a continuous distribution F(x) crosses the vertical steps of its empirical distribution $F_n(x)$. $F_n(x)$ is based on n independent observations from a distribution F(x). Smirnov showed that $P(V_n \leq t\sqrt{n}) \to 1 - e^{-t^2/2}$ for $t \geq 0$ and $P(V_n \leq t\sqrt{n}) \to 0$ for $t < 0$ as $n \to \infty$, independent of F(x). By generalizing this result to counting the number of crossings of $F(x) + \lambda/\sqrt{n}$ with $F_n(x)$, and $F(x) \pm \lambda/\sqrt{n}$ with $F_n(x)$, Smirnov was able to show that $P(\sup F_n(x) - F(x) \leq \lambda/\sqrt{n}) \to 1 - e^{-2\lambda^2}$ and $P(\sup | F_n(x) - F(x) | \leq \lambda/\sqrt{n}) \to \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2\lambda^2}$ for $\lambda \geq 0$, as $n \to \infty$. The purpose of this paper is to derive the exact distribution of V_n and define a goodness-of-fit test based on V_n. Actually we consider C_n, which is the number of crossings of F(x) with the horizontal steps of $F_n(x)$. Note that $C_n = V_n - 1$. The reason for choosing C_n is that there is a simple relationship between C_n and the transformed observations $F(X_i)$, which allows the test to be applied without recourse to actually plotting F(x) and $F_n(x)$ and counting the number of crossings. The procedure is described in section 4.

2. THE DISTRIBUTION OF C_n

We disregard the steps $F_n(x) = 0$ and $F_n(x) = 1$ and consider the n - 1 steps corresponding to $F_n(x) = i/n$ for $i = 1, 2, \ldots, n - 1$ and seek $P(C_n = k)$ for $k = 0, 1, \ldots, n - 1$. If there are exactly i observations less than $F^{-1}(i/n)$, then the $i^{\underline{th}}$ step will cross F(x). Conversely, if the $i^{\underline{th}}$ step crosses, there

must have been exactly i observations less than $F^{-1}(i/n)$. Since $P(X \leq F^{-1}(i/n)) = P(F(X) \leq i/n) = i/n$, we can use the binomial distribution to get

$$P(E_i) = \binom{n}{i}\left(\frac{i}{n}\right)^{i}\left(1 - \frac{i}{n}\right)^{n-i} \qquad i = 1, 2, \ldots, n - 1, \tag{2.1}$$

where E_i is defined as the event that the $i^{\underline{th}}$ step crosses $F(x)$. For $1 \leq i < j \leq n - 1$, we compute the simultaneous occurrence of E_i and E_j by conditioning on E_j as

$$P(E_i E_j) = P(E_i \mid E_j)\ P\ (E_j).$$

The $i^{\underline{th}}$ step will cross, given that E_j has occurred, if and only if i of the j observations in $(0, j/n)$ are in $(0, i/n)$, i.e.,

$$P(E_i E_j) = \left\{\binom{j}{i}\left(\frac{i}{j}\right)^{i}\left(1-\frac{i}{j}\right)^{j-i}\right\}\left\{\binom{n}{j}\left(\frac{j}{n}\right)^{j}\left(1-\frac{j}{n}\right)^{n-j}\right\}. \tag{2.2}$$

Using the notation $P(E_i) = {}_nP_i$ and $P(E_i E_j) = {}_nP_{ij}$ we have $P(E_i E_j) = ({}_jP_i)\ ({}_nP_j)$. Extending this notation we get, for $1 \leq i < j < k \leq n-1$,

$$\begin{aligned} {}_nP_{ijk} = P(E_i E_j E_k) &= P(E_i E_j \mid E_k)\ P(E_k) \\ &= ({}_kP_{ij})\ ({}_nP_k) \\ &= ({}_jP_i)\ ({}_kP_j)\ ({}_nP_k) \end{aligned} \tag{2.3}$$

This scheme can be extended to the simultaneous occurrence of any k events E_i, $k = 1, 2, \ldots, n - 1$. We define

$$S_j = \sum {}_nP_{i_1 i_2 \cdots i_j} \qquad j = 1, 2, \ldots, n - 1\ , \tag{2.4}$$

where the summation is taken over all subscripts $1 \le i_1 < i_2 < \ldots i_j \le n$, and the indicator variables

$$X_i = \begin{cases} 1 & \text{if the } i^{\underline{th}} \text{ step crosses} \\ 0 & \text{otherwise} \end{cases}$$

for $i = 1, 2, \ldots, n-1$. Since $s^{X_i} \equiv 1 + X_i(s-1)$, we have

$$\prod_{i=1}^{n-1} s^{X_i} = 1 + \sum X_i(s-1) + \sum X_{i_1}X_{i_2}(s-1)^2 + \ldots$$

$$+ \sum X_1X_2\ldots X_{n-1}(s-1)^{n-1} \quad ,$$

where the summations are taken over the ranges as in (2.4). Since $C_n = \sum_{i=1}^{n-1} X_i$, we have the probability generating function of C_n as

$$P(s) = E\left(\prod_{i=1}^{n-1} s^{X_i}\right) = 1 + \sum_{j=1}^{n-1} (s-1)^j S_j \tag{2.5}$$

because $E(X_{i_1}X_{i_2}\ldots X_{i_j}) = {}_nP_{i_1i_2\ldots i_j}$.

Note that we can write

$${}_nP_i = \frac{n!}{n^n}\,\frac{i^i}{i!}\,\frac{(n-i)^{n-i}}{(n-i)!} \quad ,$$

and, in general,

$${}_nP_{i_1i_2\ldots i_j} = \frac{n!}{n^n}\,\frac{i_1^{\,i_1}}{i_1!}\,\frac{(i_2-i_1)^{i_2-i_1}}{(i_2-i_1)!}\ \cdots\ \frac{(n-i_j)^{n-i_j}}{(n-i_j)!} \quad .$$

Now

$$S_j = \sum {}_nP_{i_1i_2\ldots i_j} = \frac{n!}{n^n} \sum \left\{\prod_{i=1}^{j+1} \frac{\lambda_i^{\lambda_i}}{\lambda_i!}\right\} , \tag{2.6}$$

where this summation is taken over $\sum_{i=1}^{j+1}\lambda_i = n$ and $\lambda_i \ge 1$.

We calculate the unconstrained sum

$$\sum_{\lambda_i \geq 1} x^{\lambda_1 + \ldots + \lambda_j} \left\{ \prod_{i=1}^{j+1} \frac{\lambda_i^{\lambda_i}}{\lambda_i!} \right\} \tag{2.7}$$

and use the coefficient of x^n to get S_{jn}.

Using the Residue Theorem, we have

$$\frac{\lambda^\lambda}{\lambda!} = \frac{1}{2\pi i} \oint \frac{e^{\lambda z}}{z^{\lambda+1}} dz \ , \tag{2.8}$$

and

$$\sum_{\lambda=1}^{\infty} \frac{x^\lambda \lambda^\lambda}{\lambda!} = \frac{1}{2\pi i} \oint \sum_{\lambda=1}^{\infty} \frac{1}{z} \left(\frac{xe^z}{z} \right)^\lambda dz$$

$$= \frac{1}{2\pi i} \oint \frac{1}{z} \frac{xe^z}{(z - xe^z)} dz$$

$$= \frac{z(x)}{1 - z(x)} \ , \tag{2.9}$$

where $z(x)$ is the solution to $z = x\ e^z$ such that $\frac{z(x)}{x} \to 1$ as $x \to 0$. The coefficient of x^n is obtained from

$$\frac{1}{2\pi i} \oint \frac{1}{x^{n+1}} \left(\frac{z(x)}{1 - z(x)} \right)^{j+1} dx \ ; \tag{2.10}$$

and changing the variable of integration from z to x, we have

$$S_j = \frac{n!}{n^n} \frac{1}{2\pi i} \oint \frac{e^{nz}}{z^n} \left(\frac{z}{1-z} \right)^j dz = \frac{n!}{n^n (j-1)!} \sum_{i=0}^{n-j-1} \frac{n^i}{i!} \frac{(n-2-i)!}{(n-j-1-i)!} \ . \tag{2.11}$$

Notice that $S_{jn} = 0$ for $j \geq n$ in (2.4) and (2.11). Thus from (2.5)

we have

$$P(s) = 1 + \sum_{j=1}^{\infty} (s-1)^j S_j$$

$$= 1 + \frac{n!}{n^n} \frac{1}{2\pi i} \oint \frac{e^{nz}}{z^n} \sum_{j=1}^{\infty} (s-1)^j \left(\frac{z}{1-z}\right)^j dz$$

$$= 1 + \frac{n!}{n^n} \frac{(s-1)}{2\pi i} \oint \frac{e^{nz}}{z^{n-1}} \frac{1}{1-sz} \, dz \ . \tag{2.12}$$

We have

$$C(s) = \frac{P(s)}{1-s} = \sum_{k=0}^{n-1} P(C_n \leq k) s^k$$

$$= \frac{1}{1-s} - \frac{n!}{n^n} \frac{1}{2\pi i} \oint \frac{e^{nz}}{z^{n-1}} \frac{1}{1-sz} \, dz \ ,$$

which is the generating function of the cumulative probabilities. Since $C^{(k)}(0) = P(C_n \leq k)k!$, we have

$$P(C_n \leq k) = 1 - \frac{n!}{n^n} \frac{1}{2\pi i} \oint \frac{e^{nz}}{z^{n-1-k}} \, dz$$

$$= 1 - \frac{n!}{(n-2-k)!\,n^{2+k}} \ , \tag{2.13}$$

and

$$P(C_n = k) = \frac{(1+k)n!}{n^{2+k}(n-1-k)!} \ . \tag{2.14}$$

Using Stirling's formula and $\ln(1-x) \approx -x - x^2/2$ for small x, we have, for $k = t\sqrt{n}$,

$$P(C_n \le k) \approx 1 - \left(\frac{n}{n-(k+2)}\right)^{1/2} \left(\frac{n}{n-(k+2)}\right)^{n-(k+2)} \exp\left(-(k+2)\right)$$

$$= 1 - \exp\left(-\frac{1}{2} \ln\left(1 - \frac{k+2}{n}\right) - \left(n-(k+2)\right) \ln\left(1 - \frac{k+2}{n}\right) - (k+2)\right)$$

$$\approx 1 - \exp\left(-\frac{1}{2n} (k+1)(k+2)\right) \to 1 - \exp\left(-\frac{1}{2} t^2\right),$$

which is Smirnov's result.

3. RELATED RESULTS

The characteristic function of C_n may be obtained from (2.12) by substituting $s = e^{it}$ and noting that $S_{jn} = 0$ for $j \ge n$. We get

$$\Phi_n(t) = \sum_{j=0}^{n-1} \left(e^{it} - 1\right)^j S_u \; , \tag{3.1}$$

and calculate the central moments as

$$E(C_n^r) = \frac{\Phi^{(r)}(0)}{i^r} = \sum_{j=0}^{n-1} \left(\sum_{k=0}^{j} \binom{j}{k} (-1)^{j-k} k^r \right) S_{jn} \, , \tag{3.2}$$

where we can use the known result

$$\sum_{k=0}^{j} \binom{j}{k} (-1)^{j-k} k^r = \begin{cases} 0 & \text{if } r < j \\ j & \text{if } r = j \end{cases}. \tag{3.3}$$

In general, we get the recursion relation

$$S_k = \sum_{j=k}^{n-1} {}_nP_j \; S_{k-1} \qquad k = 1, 2, 3, \ldots, n-1 , \tag{3.4}$$

where $S_0 = 1$ for all n.

The characteristic function of C_n may be calculated from

$$\Phi_n(t) = E\left(e^{itC_n}\right) = \sum_{m=0}^{n-1} \sum_{k=0}^{n-1-m} \binom{m+k}{m} (-1)^k e^{itm} S_{m+k}; \tag{3.5}$$

and by a change of variable $u = m+k$, and a change in the order of summation we get

$$\Phi_n(t) = \sum_{u=0}^{n-1} \left(e^{it}-1\right)^u S_u . \tag{3.6}$$

Using (3.4), we get the form

$$\Phi_n(t) = 1 + (e^{it}-1) \sum_{j=1}^{n-1} {}_nP_j \; \Phi_j(t). \tag{3.7}$$

From (3.6), the mean and variance of C_n are

$$E(C_n) = S_1 \tag{3.8}$$

$$\mathrm{Var}(C_n) = S_1(1-S_1) + 2S_2 ,$$

and moments may be calculated using (3.5) to get

$$E(C_n^r) = \frac{\Phi^{(r)}(0)}{i^r} = \sum_{u=0}^{n-1} \left(\sum_{m=0}^{u} \binom{u}{m} (-1)^{u-m} m^r \right) S_u . \tag{3.9}$$

A table of the exact distribution of C_n up to $n = 30$ is given in table 1. In some goodness-of-fit tests, convergence to the asymptotic distribution is rapid and the asymptotic distribution may be used for small sample sizes. In the case of the Cramer-Von Mises statistic, Marshall [2] has shown that the asymptotic distribution may be used for sample sizes of three or four. In the case of C_n, convergence is slow and it is not until $n > 100$ that the exact and asymptotic distributions become reasonably close. In fact, calculating

$$P(C_n + 1 \leq \sqrt{n}\, t) = P(C_n \leq \sqrt{n}\, t - 1) = P(C_n \leq k)$$

for $k = 0, 1, 2, \ldots, n-1$, and comparing these values with $1 - e^{-t^2/2}$, where $t = \frac{(k+1)^2}{n}$ for $k = 0, 1, \ldots, n-1$, we find the maximum difference between the exact and asymptotic cumulative distributions decreases as follows:

n	=	2	3	4	5	10	20	30
Maximum difference	=	0.3679	0.2912	0.2315	0.2146	0.1469	0.1013	0.0822
n	=	40	50	60	100			
Maximum difference	=	0.0708	0.0632	0.0575	0.0443			

TABLE 1

CUMULATIVE PROBABILITIES[a]

$P(C_n \leq k)$

Sample size n	k = 0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
2	5000	1																		
3	3333	7777	1																	
4	2500	6250	9062	1																
5	2000	5200	8080	9616	1															
6	1666	4444	7222	9074	9846	1														
7	1429	3877	6501	8500	9571	9939	1													
8	1250	3437	5898	7949	9230	9807	9976	1												
9	1111	3086	5390	7439	8862	9620	9916	9991	1											
10	1000	2800	4960	6976	8488	9395	9818	9964	9996	1										
11	0909	2562	4591	6558	8122	9146	9690	9915	9985	9999	1									
12	0833	2361	4271	6181	7772	8886	9536	9845	9961	9994	9999	1								
13	0769	2189	3992	5840	7440	8622	9364	9755	9925	9983	9997	9999	1							
14	0714	2041	3746	5533	7128	8359	9179	9648	9874	9964	9992	9999	9999	1						
15	0666	1911	3529	5255	6836	8102	8988	9528	9811	9937	9983	9997	9999	9999	1					
16	0625	1797	3335	5001	6563	7852	8792	9396	9736	9901	9969	9992	9999	9999	9999	1				
17	0588	1695	3161	4770	6308	7611	8594	9256	9650	9856	9949	9985	9996	9999	9999	9999	1			
18	0555	1605	3004	4559	6070	7380	8399	9111	9555	9802	9923	9974	9993	9998	9999	9999	9999	1		
19	0526	1523	2862	4365	5848	7159	8206	8961	9453	9741	9891	9960	9987	9997	9999	9999	9999	9999	1	
20	0500	1450	2732	4186	5640	6948	8016	8810	9345	9673	9853	9941	9979	9994	9998	9999	9999	9999	9999	1

TABLE 1 (CONT'D)

CUMULATIVE PROBABILITIES

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19[b]
21	0476	1383	2614	4021	5445	6746	7831	8657	9233	9598	9809	9918	9969	9989	9997	9999				
22	0455	1322	2506	3868	5262	6554	7650	8505	9116	9518	9759	9890	9955	9984	9995	9999				
23	0435	1267	2406	3726	5090	6371	7476	8354	8998	9434	9704	9859	9939	9976	9992	9997	9999			
24	0416	1215	2313	3594	4929	6197	7306	8204	8877	9345	9645	9823	9919	9966	9987	9996	9999			
25	0400	1168	2229	3471	4777	6031	7142	8057	8756	9254	9582	9783	9896	9954	9982	9993	9998	9999		
26	0385	1124	2148	3356	4634	5872	6984	7912	8635	9160	9515	9739	9869	9940	9975	9990	9997	9999		
27	0370	1084	2074	3249	4499	5721	6831	7770	8513	9064	9445	9692	9840	9923	9966	9986	9995	9998	9999	
28	0357	1046	2005	3147	4371	5577	6683	7631	8392	8966	9372	9641	9808	9904	9955	9981	9992	9997	9999	
29	0345	1011	1941	3052	4250	5440	6541	7495	8272	8868	9297	9588	9772	9882	9943	9975	9989	9996	9999	
30	0333	0977	1880	2963	4136	5308	6403	7362	8154	8769	9220	9532	9735	9859	9929	9967	9986	9994	9998	9999

[a] Decimal point omitted.

[b] Values larger than .9999 omitted.

4. A GOODNESS-OF-FIT TEST

We use C_n as our test statistic and note that if the hypothesized distribution $F(x)$ is correct we expect C_n to be large. Thus, acceptance regions coincide with large values of C_n. We divide the interval $(0,1)$ into n equal cells of length $1/n$. The i^{th} step of the empirical distribution function crosses $F(x)$ if and only if there are exactly i observations less than $F^{-1}(i/n)$. This occurs if and only if there are exactly i transformed observations $F(x)$ in the first i cells of length $1/n$. Thus, if we denote the cell frequencies as f_i, C_n will be equal to the number of times

$$f_1 + f_2 + \dots + f_i = i \quad \text{for } i = 1, 2, \dots, n-1. \qquad (4.1)$$

To apply the test we need simply transform the sample $X_1, X_2, \dots, X_n$ to $F(X_1), F(X_2), \dots, F(X_n)$ and count according to (4.1). Critical points and significance levels may be calculated from (2.13).

REFERENCES

[1] Smirnov, N.V., "Sur les écarts de la courbe de distribution empirique," (Russian, French summary), Rec. Math. (NS) (Mat. Sborn.) Vol. 6 (48) (1939), pp. 3-26

[2] Marshall, A.W., "Ann. Math. Stat.," Vol. 29 (1958), p. 307